




APPLIED GENOMICS SIMULATION PROJECT

De novo genome assembly and functional genomics of the *Pelobatrachus nasutus* as a phenotypic plasticity, and antimicrobial defenses model

Marco Cuscunà   [University of Bologna | marco.cuscuna@studio.unibo.it]
GitHub Repository  github.com/Markus2409/Applied_Genomics_Project.

Abstract. *Pelobatrachus nasutus*, previously known as *Megophrys nasuta* and commonly referred to as the Malayan horned frog of Southeast Asia, is an anuran of the Megophryidae family. This species represents an interesting model for investigating phenotypic plasticity and antimicrobial peptides (AMPs) for biomedical research. However, a nuclear reference genome for this species is still missing, making it an ideal candidate for de novo genome assembly and annotation. Generating a high-quality genome annotation would allow major analyses to discover the genetic basis of phenotypic plasticity and the AMPs produced by this animal.

Budget: 100,000 €

Keywords: De novo genome assembly; PacBio HiFi; Hi-C sequencing; RNA-seq; *Pelobatrachus nasutus*; phenotypic plasticity; antimicrobial peptides (AMPs); functional genomics; conservation genomics

Published: 28th August 2025

1 Introduction

The Malayan horned frog (*Pelobatrachus nasutus*) is a Southeast Asian amphibian known for its striking leaf-like camouflage, achieved through a pointed snout, eyelid projection, and characteristic textured skin. Beyond these generic morphological characteristics, it has been observed that this species exhibits notable phenotypic plasticity. In fact, individuals in fragmented and drier habitats show more pronounced tubercles and mucous glands (linked to hydration and antimicrobial defense). Additionally, a variation has been observed in body size, cranial morphology, and particularly in skin coloration, which correlates with environmental gradients: darker habitats correspond to darker pigmentation, while lighter habitats are associated with lighter body coloration (Donol *et al.* [2025]).

The presence of numerous tubercles and mucous glands also makes *P. nasutus* a promising candidate for the identification of novel antimicrobial peptides (AMPs).

2 Aim of the study

The aim of this research is to generate the first reference genome of *Pelobatrachus nasutus* using long-read sequencing technologies. The second step is to identify and annotate genes associated with skin morphology, including the analysis of allelic variants in populations from protected versus fragmented habitats. Furthermore, the study will investigate antimicrobial peptide (AMP) production, supported by reference genomes of closely related species.

3 Methods

3.1 Sample collection

To perform a *de novo* genome assembly, this project requires the collection of high molecular weight (HMW) DNA (from

blood, muscle, or liver), as well as skin samples for Hi-C and RNA sequencing. We will also sample 10 individuals (5 from protected habitats and 5 from fragmented habitats) to compare SNP variants related to skin pigmentation and tubercle morphology. To investigate antimicrobial peptides (AMPs), small skin biopsies will be taken from each individual. Total RNA will be extracted from these tissues for transcriptome sequencing (RNA-seq). This will allow us to identify AMP precursor genes expressed in the skin and integrate these data with homology-guided annotation based on the reference genome of a related species (*Leptobrachium ailaonicum*).

All samples will be preserved under field conditions using RNAlater for RNA and liquid nitrogen for DNA, ensuring high integrity for downstream sequencing. Sampling will be carried out in two contrasting habitats, the same sites where Donol *et al.* [2025] collected samples for their study on morphological variation. These two habitats are located within the same biogeographic region of Sarawak, Malaysian Borneo (Figure 1):

- **Gunung Gading National Park (protected forest):** hosts undisturbed populations of *P. nasutus*.
- **Wilmar Oil Palm Plantation (fragmented habitat):** consists of a mosaic of degraded forest patches interspersed with oil palm cultivation.

Malaysia is a Party to the Nagoya Protocol; therefore, all sampling activities must comply with its regulations. This requires obtaining prior informed consent from local authorities and park officials, securing permits for tissue collection and export of genetic material from the Malaysian government and the Sarawak Forestry Department, and establishing a benefit-sharing agreement to ensure fair distribution of benefits derived from genomic resources. For the Wilmar Oil Palm Plantation, additional permission must be requested

from the private landowner to allow access and sampling.

Ethical considerations and conservation compliance will also be prioritized:

- Skin biopsies will be minimally invasive, with the use of local anesthesia.
- All individuals will be released immediately after sampling.

These protocols ensure that the project respects local biodiversity conservation rules and remains fully aligned with the directives of the Nagoya Protocol.

Malaysia is a Party to the Nagoya Protocol; therefore, all sampling activities were carried out in compliance with its regulations. This included obtaining prior informed consent (PIC) from local authorities and park officials, securing permits for tissue collection and export of genetic material from the Malaysian government and the Sarawak Forestry Department, and establishing a benefit sharing agreement (MAT, Mutually Agreed Term) to ensure fair distribution of benefits derived from genomic resources. For the Wilmar Oil Palm Plantation, additional permission was obtained from the private landowner to allow access and sampling.

Ethical considerations and conservation compliance were also prioritized:

- Skin biopsies were minimally invasive, with the use of local anesthesia.
- All individuals were released immediately after sampling.

These protocols ensured that the project respected local biodiversity conservation rules and remained fully aligned with the directives of the Nagoya Protocol.



Figure 1. Sampling sites in Sarawak, Malaysian Borneo: Gunung Gading National Park (protected forest) and Wilmar Oil Palm Plantation (fragmented habitat).

3.2 DNA Extraction and Sequencing

To generate a chromosome-level reference genome for *P. nasutus*, we prepared a high molecular weight (HMW) DNA sample. For this purpose, we used the Qiagen Genomic-tip system, which is widely adopted for long-read sequencing because it yields pure DNA with minimal fragmentation. This approach was particularly suitable for amphibian genomes such as the one analyzed here, which are often large and repeat-rich (Kosch *et al.* [2025]).

After extraction, the quality and quantity of the DNA were assessed using a Qubit fluorometer, which provided accurate quantification of double-stranded DNA (≥ 20 ng/ μ l), and by spectrophotometric ratios ($A_{260}/_{280} \approx 1.8$ – 2.0 and $A_{260}/_{230} > 2.0$) to evaluate purity. Furthermore, the integrity of HMW DNA was checked by pulsed field gel electrophoresis to ensure fragment sizes >50 kb, as recommended for PacBio HiFi sequencing.

Finally, a small aliquot of the extracted DNA was used for a pilot sequencing run to evaluate fragment length distribution and sequencing quality. The Quality Value (QV), reported by the base calling software, served as an indicator of read accuracy. For example, a QV of 30 corresponds to 99.9% base call accuracy. This preliminary test ensured that the DNA was suitable for scaling up to full long-read sequencing.

Sequencing was performed with the PacBio HiFi platform, which provides highly accurate reads (average QV ≥ 30) with a typical length of 15–20 kb, sufficient to resolve repetitive regions common in amphibian genomes. A total coverage of approximately 30–40 \times was obtained to ensure both completeness and accuracy of the assembly.

3.3 Sequence assembly

Raw long reads were assembled using *hifiasm*, an assembler specifically optimized for PacBio HiFi data (Cheng *et al.* [2021]). The assembly was then polished to correct residual base calling errors.

To achieve a chromosome level genome, Hi-C libraries were prepared from fresh tissue of the reference individual and sequenced on an Illumina NovaSeq platform. Hi-C contact maps were used in combination with 3D-DNA (Dudchenko *et al.* [2017]) and Juicebox (Robinson *et al.* [2018]) for scaffolding and manual curation of potential mis-joins.

Assembly quality was assessed using multiple metrics:

- **BUSCO** (Tetrapoda/Anura dataset): to evaluate gene completeness (Tegenfeldt *et al.* [2025]).
- **QV** (k-mer based): to measure consensus accuracy (Rhie *et al.* [2020]).
- **N50 and LAI** (LTR Assembly Index): to assess contiguity and repeat assembly quality (Ou *et al.* [2018]).

The final assembly resulted in a high quality, chromosome-level reference genome of *P. nasutus*, suitable for downstream annotation and comparative analyses.

3.4 Functional Annotation

The assembled genome of *P. nasutus* was annotated through a combination of *ab initio* prediction, RNA-seq, and homology based approaches. Amphibian genomes, especially

those of anurans, are particularly rich in repetitive elements, including LINEs, LTRs, and transposons (Kosch *et al.* [2025]). For this reason, the first step was to identify *de novo* the repetitive elements with RepeatModeler2 (Flynn *et al.* [2020]) and mask them using RepeatMasker (Smit *et al.* [2015]) to prevent spurious gene predictions.

After that, RNA-seq was performed directly on skin tissue samples from the same individual ($RIN \geq 7$), with the aim of focusing the annotation and confirming the synthesis of AMPs. Total RNA was extracted using the Qiagen RNeasy kit, and stranded poly(A)+ libraries were prepared from 2–3 biological replicates. Libraries were sequenced on an Illumina NovaSeq platform. Reads were quality filtered and mapped to the assembled genome using STAR (Dobin *et al.* [2013]). Gene prediction was performed with BRAKER2 (Gabriel *et al.* [2024]), which combines RNA-seq evidence with AUGUSTUS (Stanke *et al.* [2004]) *ab initio* models to improve exon–intron boundary prediction. To generate a high-confidence gene set, the results from BRAKER2 were refined using the MAKER pipeline (Cantarel *et al.* [2008]), integrating homology from related amphibian genomes such as *Leptobranchium ailaonicum*, whose genome has already been studied (Li *et al.* [2019]) (Figure 2).

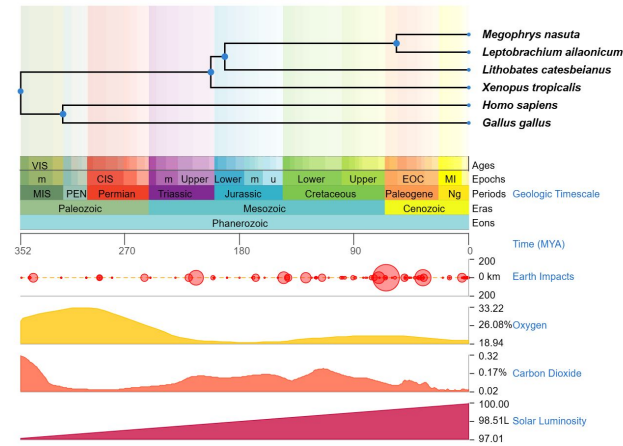


Figure 2. TimeTree-based phylogeny showing the evolutionary placement of *Pelobatrachus* (*Megophrys*) *nasutus* relative to the closely related *Leptobranchium* (*Vibrissaphora*) *ailaonicum* (both Megophryidae) and to other representative vertebrates. Divergence times are displayed against the geologic timescale. Data source: TimeTree database (Kumar *et al.* [2017]).

Special attention was given to gene families of ecological and biomedical interest such as antimicrobial peptides (AMPs), where candidate precursors were identified from the skin transcriptome, confirmed by RNA-seq expression, and validated through homology searches (BLAST, Altschul *et al.* [1990] and HMMER, Hancock and Bishop [2004]). The resulting peptides were then compared against the APD3 database (Antimicrobial Peptide Database, Wang *et al.* [2016]). Genes associated with skin morphology (involved in the synthesis of keratins, extracellular matrix proteins, and mucins) were also annotated through targeted homology searches and retained for downstream SNP analysis of populations from protected versus fragmented habitats.

The completeness of the final annotation set was assessed using BUSCO (Tetrapoda/Anura dataset) and compared with published amphibian genomes to ensure biological plausibility.

3.5 Population resequencing and SNP analysis

The second part of the analysis focused on a preliminary but promising comparison between populations sampled from protected and fragmented environments, with the aim of annotating and discovering SNPs potentially involved in the phenotypic differences observed between habitats Donol *et al.* [2025]. To investigate population-level variation and phenotypic plasticity, skin tissue samples collected from ten individuals (five from a protected forest and five from fragmented habitats) were used. Genomic DNA was extracted and sequenced at $15\times$ coverage per individual using Illumina NovaSeq PE150. Reads were quality filtered and mapped against the previously assembled chromosome-level reference genome of *P. nasutus* using BWA-MEM2 (Md *et al.* [2019]), a fast and accurate short-read aligner optimized for Illumina data, with PCR duplicates removed. Variant calling was performed with the Genome Analysis Toolkit (GATK, der Auwera *et al.* [2002]), which reconstructs local haplotypes and outputs SNP and indel calls. Since resequencing was performed at medium coverage ($15\times$), genotypes were called with some uncertainty. To address this, population genetic analysis was based on genotype likelihoods rather than hard genotype calls. For this purpose, we implemented ANGSD (Analysis of Next Generation Sequencing Data) (Korneliusson *et al.* [2014]), which estimates allele frequencies while accounting for variable sequencing depth. Principal component analysis (PCA) was performed directly on genotype likelihoods using PCAngsd (Meisner and Albrechtsen [2018]), a toolkit optimized for low-coverage genomic data. As expected, the two populations are predicted to cluster according to their environment (protected vs. fragmented).

To further assess genomic differentiation between populations, Wright’s fixation index F_{ST} (1) was calculated using ANGSD. The fixation index is defined as:

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad (1)$$

where H_T represents the total expected heterozygosity across populations, and H_S is the average expected heterozygosity within subpopulations. High F_{ST} values indicate strong genetic differentiation (Wright [1978]).

Finally, environmental association tests were conducted using BayPass (Gauthier [2025]), a Bayesian framework that detects correlations between allele frequencies and ecological variables (here, protected vs. fragmented habitats). Among the BayPass outputs, SNPs located near genes associated with skin morphology (identified in the upstream annotation analysis) were highlighted as relevant signals of local adaptation.

4 Budget

The study was performed with a total budget of approximately 100,000 €. The entire pipeline was designed to remain within the budget, balancing quality of results with financial constraints. A report of the estimated costs is provided in Table 1, which outlines the main categories of expenditure including fieldwork and permits, DNA and RNA extraction, sequencing and computational resources.

Table 1. Estimated budget for the *P. nasutus* genome project.

| Item | Estimated Cost (€) | Notes |
|---|--------------------|---|
| Fieldwork & permits | 4,800 | Sampling in Gunung Gading NP & Wilmar plantation; ABS/Nagoya compliance |
| DNA extraction (HMW + resequencing samples) | 3,250 | Qiagen Genomic-tip |
| RNA extraction (skin only) | 1,420 | RNAlater + Qiagen RNeasy kits |
| PacBio HiFi sequencing | 34,780 | 30–40× coverage, 1 HQ individual |
| Hi-C sequencing | 8,230 | Hi-C libraries from fresh tissue, sequenced on Illumina NovaSeq PE150 |
| Illumina resequencing (10 individuals, 15×) | 14,650 | Illumina NovaSeq PE150 |
| RNA-seq (skin samples, 2–3 replicates) | 6,180 | Illumina NovaSeq PE150 |
| Computational costs & Cloud storage | 11,740 | Assembly, annotation, population genomics analyses |
| Consumables & contingency | 8,560 | Tubes, barcoding, dry shipper |
| TOTAL | 93,610 | Within the 100,000 € budget |

5 Expected Results

We expect to generate the first high-quality reference genome of *Pelobatrachus nasutus* at chromosome level resolution, with high contiguity ($N50 > 20$ Mb) and completeness (BUSCO $> 90\%$). In addition, we expect to obtaining a well annotated genome, guided by skin RNA-seq data and homology with *Leptobrachium ailaonicum*, with particular focus on antimicrobial peptide (AMP) precursors and genes associated with skin morphology (keratins, mucins, extracellular matrix proteins).

We further expect a moderate genetic differentiation between populations from protected and fragmented habitats, with F_{ST} values ranging between 0.05 and 0.15, as revealed by PCA performed on outputs produced by ANGSD. Finally, we expect to identify specific SNPs associated with environmental conditions using BayPass, particularly those located near genes related to skin adaptation and antimicrobial defense, highlighting in this way possible targets of local selection and phenotypic plasticity.

6 Data Submission

All raw sequencing reads (PacBio HiFi, Hi-C, Illumina resequencing, and RNA-seq), together with the assembled and annotated genome and the VCF files, will be submitted to the European Nucleotide Archive (ENA). This ensures compliance with the FAIR principles and enables reproducibility and reuse by the scientific community.

7 Conclusions

This project reports the generation of a high-quality *P. nasutus* reference genome and positions this species as a genomic model for future studies on phenotypic plasticity. In addition, the project also provides valuable genomic resources with direct applications in biomedicine, particularly through the discovery and characterization of antimicrobial peptides (AMPs).

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215. DOI: 10.1016/S0022-2836(05)80360-2.
- Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., and Yandell, M. (2008). Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18. DOI: 10.1101/gr.6743907.
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18. DOI: 10.1038/s41592-020-01056-5.
- der Auwera, G. A. V., Carneiro, M. O., Hartl, C., Poplin, R., Angel, G. D., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Alshuler, D., Gabriel, S., and DePristo, M. A. (2002). Gatk best practices. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]*, 11.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: Ultrafast universal rna-seq aligner. *Bioinformatics*, 29. DOI: 10.1093/bioinformatics/bts635.
- Donol, C. M., Zainudin, R., and Deka, E. Q. (2025). Morphological variation of *pelobatrachus nasutus* (schlegel, 1858) (order: Anura, family: Megophryidae) from different localities in sarawak. *Journal of Sustainability Science and Management*, 20:390–406. DOI: 10.46754/jssm.2025.02.013.
- Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., and Aiden, E. L. (2017). De novo assembly of the *aedes aegypti* genome using hi-c yields chromosome-length scaffolds. *Science*, 356. DOI: 10.1126/science.aal3327.
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., and Smit, A. F. (2020). Repeatmod-

- eler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, 117. DOI: 10.1073/pnas.1921046117.
- Gabriel, L., Brúna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., and Stanke, M. (2024). Braker3: Fully automated genome annotation using rna-seq and protein evidence with genemark-ets, augustus, and tsebra. *Genome Research*, 34:769–777. DOI: 10.1101/gr.278090.123.
- Gauthier (2025). Baypass software for population genomics. <http://www1.montpellier.inra.fr/CBGP/software/baypass/>. Accessed: 2025-08-25.
- Hancock, J. and Bishop, M. (2004). hmmer. *Dictionary of Bioinformatics and Computational Biology*. DOI: 10.1002/9780471650126.dob0323.pub2.
- Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). Angsd: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15. DOI: 10.1186/s12859-014-0356-4.
- Kosch, T. A., Crawford, A. J., Mueller, R. L., Valero, K. C. W., Power, M. L., Rodríguez, A., O’Connell, L. A., Young, N. D., and Skerratt, L. F. (2025). Comparative analysis of amphibian genomes: An emerging resource for basic and applied research. *Molecular Ecology Resources*, 25. DOI: 10.1111/1755-0998.14025.
- Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). Timetree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, 34. DOI: 10.1093/MOLBEV/MSX116.
- Li, Y., Ren, Y., Zhang, D., Jiang, H., Wang, Z., Li, X., and Rao, D. (2019). Chromosome-level assembly of the mustache toad genome using third-generation dna sequencing and hi-c analysis. *GigaScience*, 8. DOI: 10.1093/giga-science/giz114.
- Md, V., Misra, S., Li, H., and Aluru, S. (2019). Efficient architecture-aware acceleration of bwa-mem for multicore systems. In *Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019*. DOI: 10.1109/IPDPS.2019.00041.
- Meisner, J. and Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth ngs data. *Genetics*, 210. DOI: 10.1534/genetics.118.301336.
- Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the ltr assembly index (lai). *Nucleic acids research*, 46. DOI: 10.1093/nar/gky730.
- Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, 21. DOI: 10.1186/s13059-020-02134-9.
- Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., and Aiden, E. L. (2018). Juicebox.js provides a cloud-based visualization system for hi-c data. *Cell Systems*, 6. DOI: 10.1016/j.cels.2018.01.001.
- Smit, A., Hubley, R., and Grenn, P. (2015). Repeatmasker open-4.0.
- Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). Augustus: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32. DOI: 10.1093/nar/gkh379.
- Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E. M., and Kriventseva, E. V. (2025). Orthodb and busco update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research*, 53:D516–D522. DOI: 10.1093/nar/gkae987.
- Wang, G., Li, X., and Wang, Z. (2016). Apd3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44. DOI: 10.1093/nar/gkv1278.
- Wright, S. (1978). Evolution and the genetics of populations. chicago, university of chicago press. *University of Chicago Press*, 4.